Towards the development of a computerised method of analysing and estimating the vocabulary size of ESL students

Ian Smallwood

INTRODUCTION

I describe the reasoning behind, and initial stages of, research into the development of a method of analysing and estimating the probable size of an individual's lexicon — both what are often termed the 'productive' and the 'receptive' lexicons (Nation 1982:19). I will also give an outline of future research into the analytic techniques employed.

SOME PROBLEMS OF VOCABULARY ANALYSIS

The general approach adopted in this proposed method of vocabulary measurement follows, to some extent, the models used by early researchers such as Seashore & Eckerson (1940) and Diack (1975) as well as more recent models such as that used by Goulden et al. (1990), in that it assumes that knowledge of a subset of words represents knowledge of a larger set of words. For example, knowledge of a subset of ten words could be taken to represent knowledge of a set of 1,000 words. In this case a subject’s knowledge of ten subsets would indicate a total vocabulary size of 10,000 words.

Although this approach has been widely adopted by researchers, there does not seem to be much agreement as to how the basic problems associated with estimating mental lexicons can best be overcome. More specifically, there is little agreement on what approach can best address the questions of:

1. which words should be counted;
2. which words can be taken as being represented by the subset of words;
3. whether the analysis should focus on the ‘breadth’ of the lexicon or its ‘depth’ (i.e. a count of how many words can be recognised/used or how thoroughly a word family has been learned);
4. what differences there are between ‘productive’ and ‘receptive’ lexicons;
5. what the best way is of determining whether a word is ‘known’.

WHICH WORDS SHOULD BE COUNTED?

One of the major causes of wide discrepancies in vocabulary estimates of the same or similar groups of individuals — for example a range of 3,000 to 216,000 words (Diack 1975; Diller 1978; cf. Dodd 1993) — is the definition of what actually constitutes a ‘word’. Lower estimates tend to exclude all derivatives and proper nouns and only count lemmata, whereas high estimates tend to count all words, including derivatives, inflections and proper nouns. Goulden et al. (1990:344) suggest that the total number of possible words to be learned and analysed should be kept to a minimum by only including lemmata (or ‘base words’) in the analysis, the assumption being that regular derivative constructions such as prefixes and suffixes can easily be learned and so ‘if a learner knows govern and is familiar with the prefix mis-, then misgovern requires little if any additional learning’ (ibid.). This may, perhaps, have some truth if the analysis is restricted to native speakers, but it is often not the case when non-native speakers are also included. Take, for example, the words sing or take and the prefix mis to form the words missing and mistake, which are clearly not words comprising the lemmata ‘sing’ and ‘take’ with the prefix mis. The approach adopted here is to exclude regular plurals and regularly inflected verb forms, but to include all other derivatives which appear on the word lists. With some exceptions (discussed later), nouns which appear in the word lists would also be included.

WORD FREQUENCY LISTS

The question of which words to include in the count, of course, directly affects the way in which the count can be made. Almost all studies of the size of lexicons, particularly receptive lexicons, have been based on either dictionaries or a published general word frequency list (e.g. Thorndike & Lorge 1944; West 1953; Kučera & Francis 1967). Both of these approaches are likely to lead to inaccurate counts. The use of a dictionary tends to inflate the number of possible words as very high frequency ‘function’ words are given the same status as rarely occurring and very complex technical terms. Inaccuracies can also occur as a result of variations in which derivative forms of a word are included or excluded. It would also
seem logical that an accurate formula for multiplying the subset should not be based simply on a list of all possible words in the language, but also on some ranking of the words, either in terms of frequency or of difficulty. This is the reason behind the use of word lists which indicate the frequency with which the words occur. However, the frequency counts used in research so far have tended to be frequency counts of general, and often out-dated, usage and so do not necessarily reflect the exposure the learner has had to the words. This is again particularly the case for L2 students when the frequency counts were made using authentic L1 texts.

In order to overcome this inaccuracy, it is necessary to construct a word frequency count from a corpus of text which the target users of the lexicon test are likely to have been exposed to. For native speakers of the language this would be an almost impossible task. However, for L2 students it is, theoretically, a much easier undertaking. The project outlined in this paper involves the construction of a corpus of most of the English-language textbooks published in Hong Kong from Primary level to Upper Secondary level, including ELT materials, subject-based materials and testing materials. As the majority of learners in Hong Kong appear not to be influenced to any significant degree by the authentic language input available to them via television, newspapers, and radio, it would seem to be a reasonable assumption that the corpus would be an accurate reflection of the language that the Hong Kong learners had been exposed to prior to the lexicon test. Therefore, a frequency list constructed from this corpus seems likely to provide accurate information when estimating the lexicons of the learners in Hong Kong.

These frequency lists could be further refined by making lists from different sections of the corpus. For example, if the learner had followed a Science programme through school, the frequency list could take account of this; if the learner left school after form 5, the list could reflect this by only analysing texts presented up to the point when the learner stopped studying. If necessary other texts, such as a corpus of the local English-language newspapers, which is readily available, could also be included. The problem of the range (McCarthy 1990:69) of use of the word can also be answered here by checking whether the word is a very high frequency word in a specific genre of text and low in all the rest. The frequency lists constructed according to these principles would then be divided into separate word lists, each containing the number of words necessary for the level of accuracy required in the estimate (i.e. the lower the number of words contained in each word list, the more accurate the estimate). If the estimated total vocabulary size is likely to be small, then the word lists could also be small, to give a more precise estimate.
A further advantage of this approach to the formation of word lists is that questions of which words to include in the count could more readily be answered. Some words, such as regular plurals and verb forms, could be deleted either manually or automatically. The commoner derivational forms of a word will appear in earlier subsets, the less common (and it is assumed that ‘less common’ probably means less likely to be known) would appear in later subsets. A learner may know the commoner forms of a word but not necessarily the less common derivatives. Proper nouns which are romanised transcriptions of Chinese names, for example, could be deleted, but the names of countries, which often cause great difficulty to a learner, could be retained. In short, the word lists could be modified to suit the specific purposes and preferences of the investigator.

**BREADTH OR DEPTH?**

It is also possible that the question of whether to analyse the breadth of the lexicon or its depth could also be answered to some extent by adopting this approach. The breadth of the lexicon — the total number of words likely to be known (Anderson & Freebody 1981) — is relatively easy to estimate using this approach. The depth of understanding is often estimated by the number of derivatives of a lemma which are known. In many cases, the derivatives are more commonly used than the lemma, therefore this approach can be rejected. A more appropriate method seems to be to tabulate word families with an indication of where on the frequency ranges the word types within a family occur. If, for example, a testee has an estimated total lexicon of 2,500 words, the tabulated word families will show the number of derivatives of each lemma which are estimated to be known. In brief, the actual composition of the breadth of vocabulary could be calculated in terms of word families and not only word types, and a relatively simple formula could be calculated to give a numerical value to the ‘depth’ of the lexicon.

A recurring source of inaccuracy in word counts is the problem of homographs or polysemous words, and semantic units of more than one word, for example verbal/prepositional phrases. As there is a high semantic component involved in recognising these items, any computational analysis cannot guarantee total accuracy. However, by including phrases in the word lists as single entries and including, wherever possible, determining collocates with homographs, the problem can, perhaps, be reduced. Of course, the use of word lists as a base for the analysis precludes any extensive contextualisation of vocabulary, which means the words which are assumed to be known may, in fact, have been misused. However, as Read (1993:357) points out, analysing ‘words in isolation may be the only practical way of achieving the necessary coverage’.
HOW CAN VOCABULARY KNOWLEDGE BE MEASURED?

A final area of disagreement among vocabulary researchers is how the knowledge of a word is to be assessed. Can a word be assumed to be known if it can simply be recognised as belonging to the total L2 lexicon (as in the ‘Lingua Vocabulary Tests’, Meara 1993); if it can be recognised and the semantic element of the word understood; or does it have to be produced in a meaningful way before we can accept that it is fully ‘acquired’? The first option was rejected as it was felt that, ‘although in certain contexts verbal report and self-ratings by test-takers are practical and valid methods of assessing vocabulary knowledge, in most testing situations it is clearly inappropriate to rely on unverified self-report’ (Read 1993:356).

Nation (1982:19) differentiates between the latter two definitions of knowing a word by referring to them as ‘receptive’ and ‘productive’ lexicons respectively. There seems to be little doubt that there will be a significant difference between the two, but both are of equal interest. The approach being developed here can be adapted to estimate the size of either.

Clearly, any approach which attempts to estimate ‘productive’ vocabulary (i.e. words which can be used meaningfully in context) must have as large a sample of language as possible for analysis. In this case the language output must be written, as voice recognition software has not been sufficiently developed at present. The sample must also be in electronic format for the computerised analysis — this should present an ever-decreasing problem as more and more learners use word-processing programs and as text-scanners become more sophisticated. The sample body of text can be matched against the word lists, in sequence, beginning with the most commonly used words, and the program can determine the point at which insufficient words of a subset are known. As any frequency count of a corpus will have a very large number of words which only occur once, it is impossible to sequence all of the words in the corpus accurately. The final word list will contain all these single occurrences of words and a different ratio/formula can be calculated for words the subject uses from this list. As all words in this list are words only occurring once in almost the entire language exposure the subject has had, the number of items used and known is likely to be very small. Any spellings are regarded as non-words and are not accepted, as they do not appear on any word list. This program has, to some extent, already been developed and is described more fully below.

The analysis of receptive vocabulary can also be done computationally. Each subset of an individual word list will randomly present a given number of words in sequence. The testee keys in either a synonym in English or a translation in the pre-identified L1 and the program monitors correct and incorrect answers. It was
decided to provide the option of an L2 synonym or a translation as it was felt, contrary to the definition of receptive learning given by Nation (1982:19) as being limited to the ability to ‘recall the translation of the foreign word’, that receptive ‘use’ of a foreign word does not always require a translation. The testee is told to enter multiple answers if more than one meaning of the word is known. All correct answers for polysemous words are logged in the appropriate word list (i.e. one meaning may be in word list 5 and the other in word list 8, for example). The program automatically stops when the threshold word list is reached. For both the receptive test and the productive analysis, further research is required to ascertain what percentage of a subset represents knowledge of the subset as a whole.

A PRELIMINARY ANALYSIS

In an attempt to test some of the general principles outlined above, a small-scale investigation was carried out to compare the productive vocabulary of two groups of L2 students: those scoring an ‘A’ grade and those scoring an ‘F’ grade in the Hong Kong Advanced Level Use of English examination composition paper. Both of these were then compared to the vocabulary used in textbooks written for Secondary 5 students (i.e. two years below the school level of the subjects studying A-level). These samples of text were chosen in an attempt to test whether or not the analytical methods being employed could significantly differentiate between what were likely to be three different sizes of ‘productive’ lexicon, although no research could be found which examined whether vocabulary was a factor in establishing Use of English composition grades.

This study explored two related hypotheses:

1. Large productive lexicons would result in a lower total type/token ratio.
2. Large productive lexicons would result in a greater number of less frequent word types.
The text samples

Three small corpora were created, of approximately 7,500 words each. One was created from a random selection of complete texts which were awarded an ‘F’ grade at Use of English, one from complete texts which were awarded an ‘A’ grade and the remaining corpus from complete texts written as comprehension passages for a Hong Kong Certificate of Education (HKCE) examination practice papers publication. Complete texts were used as it is possible that different stages of a composition may display different lexical characteristics because of the function that stage is serving. The HKCE examination was chosen as it was felt that the passages used in ‘Use of English’ practice papers were authentic texts with little or no modification for the L2 learner and would be less representative of the level which the learners could reasonably be expected to attain. It should, however, be noted that it could reasonably be expected that Use of English passages would extend both parts 1 and 2 of the hypothesis stated above.

The word lists

As the complete corpus of materials used in Hong Kong secondary schools is not at a stage where it could be used in this analysis, a limited set of three word lists originally designed by Nation et al. (1992) was used. The first word list includes the 1,000 most frequent words of English. The second includes the second 1,000 most frequent words and the third includes words which are not found in the first 2,000 words of English but which are frequent in upper-secondary school and tertiary texts from a range of subjects. All three word lists include the base forms of words as well as the commonest derived forms and therefore the actual word count in each list is approximately 2,000. Any word beyond the 6,000 most frequent was counted but not analysed further. It is accepted that these word lists would be unlikely to give an accurate estimation of the actual sizes of the lexicons being analysed. They would need to be significantly smaller and more precisely structured to give such an accurate estimate. They would also need to be constructed from an archive more representative of the language the subjects had previously been exposed to. It was felt, however, that the word lists would be able to show significantly different trends in vocabulary frequency distribution patterns for the three samples and therefore show that further research into the analytical method being proposed was valid.
PROCEDURE

Each of the corpora was analysed individually by the program to determine the total number of word types used and the number of word types used within each of the three word lists in addition to types used which were not on any of the word lists (also given as a percentage of the total word types used by that sample). As the word lists being used are incomplete, the final category of word types, those not found on any word list, had to be checked manually and any words not accepted by this analysis discussed earlier in this paper deleted from the list. The results of each of the three analyses were then compared for significant differences.

RESULTS

The results of the three analyses are shown in Table 1. The word lists 1–3 are presented in order of word-frequency (word list 1 being the most frequent), and the column ‘Rest’ being the group of words which are words accepted by the analysis but not included in the three word lists.

Table 1: An analysis of the vocabulary frequency patterns of the three text samples.

<table>
<thead>
<tr>
<th></th>
<th>Totals</th>
<th>Word lists</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tokens</td>
<td>Types</td>
</tr>
<tr>
<td>L1 Text</td>
<td>7504</td>
<td>2031</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade ‘A’ L2 essays</td>
<td>7454</td>
<td>1452</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade ‘F’ L2 essays</td>
<td>7461</td>
<td>1097</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Analysis of results

As can be seen from the actual number of word types produced in each of the sample texts, there is a clear downward trend from L1 to ‘F’-grade texts in the likely productive lexicons of the respective subjects (L1 texts used almost double the number of types of ‘F’ grade texts). The trend can, perhaps, be generalised
somewhat by an analysis of the type-token ratios for each of the samples (given that the token count is almost the same for each group), represented in Figure 1.

![Bar graph showing type-token ratios for text samples.](image)

Figure 1: Type-token ratios for text samples.

The increase of type-token ratio from L1 to 'F'-grade texts would support the hypothesis that there is a decrease in productive lexicons from L1 to 'F'-grade texts. For the proposed analytical methods to succeed, it is necessary to establish whether this decrease is reflected in a difference in frequency distribution patterns across the word lists. In Figure 2, the decrease in productive lexicons is seen in the result that the L1 sample has the lowest percentage of types occurring in word list 1 and the highest occurring beyond the range of the word lists.
Figure 2: The word types in each word list as a percentage of the total word types for the text.

It is also interesting to note that the frequency bands which are the target of many vocabulary measurement tests (the 2,000-6,000 most frequent words in English) are the bands with least, or even no, significant differences. The major differences are much greater in the expected band of least common words, but equally significantly also in the most frequent words when taken as a percentage of the total word types produced. Further research may indicate that it is possible to compare, though clearly with less accuracy, individual total vocabulary sizes by analysing only the use of the most frequent words as a percentage of the total lexicon. This would also allow the research to avoid many of the problem areas of vocabulary measurement.
CONCLUSION

Although the word lists used for this analysis were not suitable for an accurate estimate of the productive lexicons of the three sample groups, they were suitable for an analysis of the vocabulary frequency distributions of each group. As such, the analysis did indicate that the general approach to vocabulary analysis being proposed here merits further enquiry. The areas still to be developed include:

1. the completion and analysis of the corpus of Hong Kong secondary school textbooks;
2. the creation of a range of word lists, based on this corpus, which could be used for subjects with differing language backgrounds such as those of varying age groups and specialist study groups such as students in the Science stream;
3. an investigation into what percentage of a subset of words represents knowledge of the set of words as a whole, including the final set of words which will be significantly larger than any other set, as it will consist of all words only occurring once in the corpus.

A final area of research which could be valuable would be an analysis of the correlation between word types used in the commonest 1,000–2,000 words, as a percentage of the total word types used, and total productive lexicon as estimated by matching the sample against the full set of word lists. If a high correlation were found to exist, it might be possible to simplify the whole measurement procedure.
Towards the development of a computerised method of analysing and estimating the vocabulary size of ESL students

Ian Smallwood

Reference


